Calibration tests for multivariate Gaussian forecasts

Wei, Wei ; Balabdaoui, Fadoua ; Held, Leonhard

Abstract: Forecasts by nature should take the form of probabilistic distributions. Calibration, the statistical consistency of forecast distributions and observations, is a central property of good probabilistic forecasts. Calibration of univariate forecasts has been widely discussed, and significance tests are commonly used to investigate whether a prediction model is miscalibrated. However, calibration tests for multivariate forecasts are rare. In this paper, we propose calibration tests for multivariate Gaussian forecasts based on two types of the Dawid–Sebastiani score (DSS): the multivariate DSS (mDSS) and the individual DSS (iDSS). Analytic results and simulation studies show that the tests have sufficient power to detect miscalibrated forecasts with incorrect mean or incorrect variance. But for forecasts with incorrect correlation coefficients, only the tests based on mDSS are sensitive to miscalibration. As an illustration, we apply the methodology to weekly data on Norovirus disease incidence among males and females in Germany, in 2011–2014. The results further show that tests for multivariate forecasts are useful tools and superior to univariate calibration tests for correlated multivariate forecasts.

DOI: https://doi.org/10.1016/j.jmva.2016.11.005

Calibration tests for multivariate Gaussian forecasts

Wei Wei

Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland

Fadoua Balabdaoui

CEREMADE, Universite Paris-Dauphine Paris, France
Email: fadoua@ceremade.dauphine.fr

Leonhard Held

Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland
Email: leonhard.held@uzh.ch

Abstract

Forecasts by nature should take the form of probabilistic distributions. Calibration, the statistical consistency of forecast distributions and observations, is a central property of good probabilistic forecasts. Calibration of univariate forecasts has been widely discussed, and significance tests are commonly used to investigate whether a prediction model is miscalibrated. However, calibration tests for multivariate forecasts are rare. In this paper, we propose calibration tests for multivariate Gaussian forecasts based on two types of the Dawid-Sebastiani score (DSS): the multivariate DSS (mDSS) and the individual DSS (iDSS). Analytic results and simulation studies show that the tests have sufficient power to detect miscalibrated forecasts with incorrect mean or incorrect variance. But for forecasts with incorrect correlation coefficients, only the tests based on mDSS are sensitive to miscalibration. As an illustration, we apply the methodology to weekly data on *Norovirus* disease incidence among males and females in Germany, 2011-2014. The results further show that tests for multivariate forecasts are useful tools and

superior to univariate calibration tests for correlated multivariate forecasts.

Keywords: Calibration test, Multivariate forecast, Dawid-Sebastiani score, Logarithmic score, Gaussian forecast, Uniform correlation matrix

---

1. Introduction

One main task for statistical analysis is to predict the future. In the past two decades, probabilistic predictions have become routine in applied forecasting. Probability forecasts were first commonly used for binary endpoints, and later extended to more general types of variables. They are widely discussed and applied in many areas: in weather forecasting; in economics and finance risk management; in clinical, chronic and infectious disease epidemiology; in health care management; in atmospheric science and many other areas. They usually take the form of ensemble forecasts, interval forecasts or density forecasts. Here we will focus on the density forecast which provides the most information. In the case of a binary event, the density forecast is the probability that the event will occur; in the case of continuous variable, a probabilistic forecast is the predictive density of the outcome of interest.

How to evaluate the performance of probabilistic forecasts is an essential question in forecast research. Gneiting et al. [19] contend that the goal of probabilistic forecasting is to *maximize the sharpness of the predictive distributions subject to calibration*. In this context, calibration refers to the statistical consistency between the probabilistic forecasts and the actual observations. Sharpness refers to the concentration of the predictive distributions. If the true data generator follows the predictive distribution, we say the forecast is ideal.

Interest in density forecasting has spurred the development of methods for their evaluation [1; 14]. Much of the literature in forecast evaluation focuses on the forecasting of univariate quantities or events. A comprehensive overview of probabilistic forecasting is given by Gneiting and Katzfuss [20], including discussion of diagnostic checks and methods for the evaluation of

probabilistic forecasts. The evaluation of forecasts heavily depends on the distribution of the forecasts. The methodology began with binary outcomes (whether it will rain tomorrow), later extended to categorical events, count data and continuous quantities [9; 26]. Many diagnostic tools have been developed for model evaluation and model selection. For univariate continuous forecasts, Dawid [12] and Diebold et al. [14] propose the use of the probability integral transform (PIT). For ideal forecasts, the PIT values are uniformly distributed. Therefore, a PIT histogram is typically used as a diagnostic tool. Gneiting et al. [19] propose proper scoring rules which can evaluate calibration and sharpness simultaneously. Three proper scores are commonly used: The Dawid-Sebastiani score (DSS) [13], the logarithmic score (LS) [10; 22] and the ranked probabilistic score (RPS) [6; 19]. Calibration tests have been developed to investigate whether the forecasts are miscalibrated. Held et al. [23] develop two types of calibration tests based on proper scoring rules. Alternatively, Mason et al. [30] suggest the use of the conditional exceedance probability (CEP) in a logistic regression framework to assess calibration of continuous probabilistic forecasts.

In recent years, the evaluation of multivariate forecasts came into focus with the proliferation of multivariate probabilistic forecasting. Tools have been developed for multivariate ensemble forecasts, for example, the multivariate rank histogram [21] and the band depth rank histogram [39]. However, for multivariate density forecasts, only a limited number of methods can be applied. Firstly, many tools for univariate forecasts do not apply to multivariate forecasts. The PIT approach fails in that it is not uniform even when the observation is drawn from the predictive distribution [21]. Although some alternative transforms are proposed to retain uniformity, they do not work perfectly. For example, a step-wise procedure for PIT is proposed by Diebold et al. [15], in which the univariate PIT values are computed sequentially based on the conditional cumulative distribution function (CDF). More specifically, the univariate PIT is first applied to the first component, then to

the conditional CDF of the second given the first, and so on. However, this approach depends on the order of the components. Methods based on CEP fail as well because the quantile of a multivariate forecast is not unique. In addition, most existing methods for the evaluation of multivariate forecasts encounter the issue of low efficacy for high dimensionality, which is usually referred to as the curse of dimensionality [3]. Gneiting et al. [21] give an overview on methods for diagnostic checking and recommend proper scoring rules to evaluate multivariate forecasts. Among all the proper scoring rules available, we have decided to use DSS, which is equivalent to LS under normality of the forecast. DSS is easy to compute, straightforward to interpret, and reported to be sensitive to mis-specified correlations [36]. In addition, DSS is a standardised score which avoids problems arising from components with incommensurable or incomparable magnitude, for example if one component has values between -1 and 1, while another component is in the range -100 to 100. Finally, it is possible to derive the first two moments of the DSS for ideal Gaussian forecasts, as we will show in Section 2.

In this paper, we develop calibration tests for multivariate Gaussian density forecasts based on the DSS. The structure of the paper is as follows. In Section 2 we introduce two types of the DSS and their properties for ideal Gaussian forecasts. In Section 3 we develop calibration tests based on the DSS to check calibration of multivariate predictions. We evaluate the power of the proposed tests analytically and via simulations in Section 4, where we pay particular attention to uniform correlation matrices [16]. As a practical application, in Section 5 we apply the tests to evaluate predictive models for the weekly number of reported *Norovirus* infections among males and females in Germany. We end with a discussion in Section 6. In this paper, symbols in bold face represent multivariate quantities or distributions, whereas normal symbols are univariate quantities.

2. David-Sebastiani scores for multivariate Gaussian forecasts

A scoring rule assigns a numerical score based on the predictive distribution and the realization. It can be viewed as a penalty of the statistical difference between an observation and a prediction and is usually negatively oriented, *i.e.*, the smaller, the better. A proper scoring rule ensures that quoting the true predictive distribution as forecast distribution is an optimal strategy in expectation. In this paper, we consider the David-Sebastiani and the logarithmic score, which are equivalent for a normal distribution. For a univariate observation x and a normal prediction $P = \mathcal{N}(\mu_P, \sigma_P^2)$, the logarithmic score is defined as

$$\mathrm{LS}(x, P) = \frac{1}{2} \ln \sigma_P^2 + \frac{(x - \mu_P)^2}{2\sigma_P^2}.$$

The Dawid-Sebastiani score $\mathrm{DSS}(x, P) = 2\,\mathrm{LS}(x, P)$ is the same up to the multiplicative constant of 2 and can also be used for other (non-normal) multivariate predictions.

For an m-dimensional normal prediction, there are two types of DSS: the multivariate score (mDSS) and the individual score based on each component (iDSS). Denoting the multivariate prediction as $\boldsymbol{P} = \boldsymbol{\mathcal{N}}_m(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$, the mDSS is computed as

$$\mathrm{mDSS}(\boldsymbol{x}, \boldsymbol{P}) = \ln|\boldsymbol{\Sigma}_P| + (\boldsymbol{x} - \boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_P),$$

where $|\boldsymbol{\Sigma}_P|$ denotes the determinant of $\boldsymbol{\Sigma}_P$. The component $\ln|\boldsymbol{\Sigma}_P|$ measures the sharpness and the quadratic form $Q(\boldsymbol{x}, \boldsymbol{P}) = (\boldsymbol{x} - \boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_P)$ is the standardised difference between the observation $\boldsymbol{x}$ and the predicted mean $\boldsymbol{\mu}_P$. The multivariate score, mDSS, gives a scalar value for any $\boldsymbol{x}$ and $\boldsymbol{P}$. The marginal predictive distribution for each component is also a normal distribution, $P_j = \mathcal{N}(\mu_j, \sigma_j^2)$, where $j = 1, \ldots, m$. For each of these, we can compute the individual DSS (iDSS) as $\mathrm{iDSS}(x_j, P_j) = \ln \sigma_j^2 + (x_j - \mu_j)^2 / \sigma_j^2$.

For multivariate forecasts of dimension m, there is one iDSS for each component ($\text{iDSS}_1, \ldots, \text{iDSS}_m$), while mDSS gives a scalar value. It is worth noting that the mDSS and iDSS are coherent if the components are independent, *i.e.*, the covariance matrix $\boldsymbol{\Sigma}_P$ is diagonal; then $\text{mDSS}(\boldsymbol{x}, \boldsymbol{P}) = \sum_{j=1}^{m} \text{iDSS}_j$, because the joint density equals the product of the marginal densities. When the correlation is not zero, mDSS can be viewed as a standardised combination of iDSS adjusted for the correlation structure.

An ideal forecast assumes that the data generating process $\boldsymbol{X}$ is exactly the same as the predictive distribution. Under the null hypothesis H_0: $\boldsymbol{X} \sim \boldsymbol{P} = \boldsymbol{\mathcal{N}}_m(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$, mDSS follows a χ^2-distribution with m degrees of freedom, but shifted by a constant:

$$\text{mDSS}(\boldsymbol{X}, \boldsymbol{P}) \underset{\text{H}_0}{\sim} \chi^2(m) + \ln|\boldsymbol{\Sigma}_P|. \tag{2.1}$$

Note that the expectation $\mathsf{E}_0\{\text{mDSS}(\boldsymbol{X}, \boldsymbol{P})\} = m + \ln|\boldsymbol{\Sigma}_P|$ depends on the logarithm of the determinant of the predictive covariance matrix, but the variance $\mathsf{Var}_0\{\text{mDSS}(\boldsymbol{X}, \boldsymbol{P})\} = 2m$ does not depend on the covariance matrix and is finite.

Similarly, we can obtain $\text{iDSS}(X_j, P_j) \underset{\text{H}_0}{\sim} \chi^2(1) + \ln \sigma_j^2$ with

$$\mathsf{E}_0\{\text{iDSS}(X_j, P_j)\} = 1 + \ln \sigma_j^2 \text{ and } \mathsf{Var}_0\{\text{iDSS}(X_j, P_j)\} = 2.$$

If two components X_j and X_k of $\boldsymbol{X}$ are dependent, then $\text{iDSS}(X_j, P_j)$ and $\text{iDSS}(X_k, P_k)$ are also correlated with correlation

$$\text{Corr}_0\{\text{iDSS}(X_j, P_j), \text{iDSS}(X_k, P_k)\} = \rho_{jk}^2, \tag{2.2}$$

where $\rho_{jk} = \text{Corr}(X_j, X_k)$. The proof is in Appendix A.

To summarise, under H_0, each component $\text{iDSS}(X_j, P_j)$ follows marginally a shifted χ^2-distribution and the vector $(\text{iDSS}(X_1, P_1), \ldots, \text{iDSS}(X_m, P_m))^\top$

has covariance matrix

$$\mathrm{Cov}_0\left\{\begin{pmatrix} \mathrm{iDSS}(X_1,P_1) \\ \vdots \\ \mathrm{iDSS}(X_m,P_m) \end{pmatrix}\right\} = 2\begin{pmatrix} 1 & \rho_{12}^2 & \dots & \rho_{1m}^2 \\ \rho_{12}^2 & 1 & \dots & \rho_{2m}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1m}^2 & \rho_{2m}^2 & \dots & 1 \end{pmatrix} =: 2\,\boldsymbol{\Omega},$$

where $\boldsymbol{\Omega}$ is the corresponding correlation matrix. It is easy to prove that $\boldsymbol{\Omega}$ is positive semidefinite.

3. Calibration tests

Suppose we have n independent multivariate normal forecasts $\boldsymbol{P}_i = \boldsymbol{\mathcal{N}}_m(\boldsymbol{\mu}_{P_i}, \boldsymbol{\Sigma}_{P_i})$ for the corresponding observations $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})^\top$, $i = 1, \dots, n$, and their data generators are $\boldsymbol{X}_i \sim \boldsymbol{\mathcal{N}}_m(\boldsymbol{\mu}_{X_i}, \boldsymbol{\Sigma}_{X_i})$. In a typical application, $\boldsymbol{P}_i$ is the one-step-ahead prediction at time i for the observation $\boldsymbol{x}_i$. Let us denote the diagonal vector of the covariance matrix $\boldsymbol{\Sigma}_{P_i}$ as $\boldsymbol{\sigma}^2_{P_i} = \mathrm{diag}(\boldsymbol{\Sigma}_{P_i}) = (\sigma^2_{i1}, \dots, \sigma^2_{im})^\top$.

Further assume that the matrix of cross-correlations $\boldsymbol{\rho}_{P_i}$ is the same for all forecast $\boldsymbol{P}_i$ and has elements $(\boldsymbol{\rho})_{jk} = \rho_{jk}, 1 \le j < k \le m$. Some additional discussion about this assumption is in Section 6. The null hypothesis of all calibration tests is that the forecasts are ideal, *i.e.*, H_0: $\boldsymbol{X}_i \sim \boldsymbol{P}_i$ for all $i = 1, \dots, n$.

Following the same approaches as Held et al. [23] for univariate forecasts, we extend two types of calibration tests to multivariate Gaussian forecasts: the unconditional test and the regression test. The unconditional test is "unconditional" in the sense that it averages the scores regardless of the underlying predictive distributions. The statistic is usually in the form of a standardised difference between the average over the scores and the expected average under the null hypothesis. The statistic of the unconditional test for normal predictions naturally complements Spiegelhalter's z-statistic [23; 38], taking the form

$$z_s = \{\overline{s} - \mathsf{E}_0(\overline{s})\} / \mathsf{Var}_0(\overline{s})^{1/2}, \tag{3.3}$$

where $\overline{s}$ is the average of the scores s_i. Based on the central limit theorem, we have $z_s \overset{a}{\underset{\mathrm{H}_0}{\sim}} \mathcal{N}(0,1)$, so under H_0, the statistic z_s asymptotically follows a standard normal distribution. Therefore a two-sided test can be constructed. The alternative test statistic discussed by Wei and Held [40],

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{s_i - \mathsf{E}_0(s_i)}{\sqrt{\mathsf{Var}_0(s_i)}} \tag{3.4}$$

is here equivalent to (3.3), because $\mathsf{Var}_0(s_i)$ does not depend on i for both mDSS and iDSS. Furthermore, the z-statistic (3.3) can provide hints to the reasons for forecast deficiency. A positive value of z_s indicates underdispersed forecasts with prediction intervals too narrow on average, or biased forecasts whose mean parameters are wrongly predicted. A negative value corresponds to overdispersed forecasts whose prediction intervals are too wide, see Riebler and Held [35] for an application in cancer prediction. This feature may provide clues how to further improve the forecasting model.

The idea of the regression approach is to regress the score s_i on its expectation $\mathsf{E}_0(s_i)$ under H_0,

$$s_i = a + b\,\mathsf{E}_0(s_i) + \epsilon_i, \tag{3.5}$$

where ϵ_i is a zero-mean residual error term. A test statistic based on the regression coefficients a and b can be constructed, since under the null hypothesis we have $a = a_0 = 0$ and $b = b_0 = 1$. The Wald test statistic

$$T_s = (\hat{a} - a_0, \hat{b} - b_0)\widehat{\boldsymbol{V}}^{-1}(\hat{a} - a_0, \hat{b} - b_0)^\top, \tag{3.6}$$

where $\hat{a}$ and $\hat{b}$ are the estimated coefficients and $\widehat{\boldsymbol{V}}$ is the estimated covariance matrix of $(\hat{a}, \hat{b})^\top$, follows asymptotically a $\chi^2(2)$ distribution under H_0.

3.1. Tests for mDSS

3.1.1. Test based on χ^2-distribution

For multivariate Gaussian forecasts, mDSS follows a shifted $\chi^2(mn)$ distribution under H_0, see Equation (2.1). Therefore, we can construct a statistic

Z_{mDSS}, assuming independence of mDSS_i:

$$Z_{\mathrm{mDSS}} = \sum_{i=1}^{n} (\mathrm{mDSS}_i - \ln|\boldsymbol{\Sigma}_{P_i}|) \underset{\mathrm{H}_0}{\sim} \chi^2(mn), \tag{3.7}$$

where mDSS_i is the multivariate DSS for $\boldsymbol{X}_i$ and $\boldsymbol{P}_i$. The statistic Z_{mDSS} works well even when the number of observations n is small. On the other hand, if n is large, Z_{mDSS} explodes, in which case the following unconditional test is recommended.

3.1.2. Unconditional test

When n is large, we apply the normal approximation by applying the central limit theorem and replace Z_{mDSS} by

$$z_{\mathrm{mDSS}} = \frac{\sqrt{n}(Z_{\mathrm{mDSS}}/n - m)}{\sqrt{2m}},$$

which is equivalent to Equation (3.3) with $\overline{s} = \overline{\mathrm{mDSS}} = 1/n \sum_{i=1}^{n} \mathrm{mDSS}_i$, *i.e.*,

$$z_{\mathrm{mDSS}} = \frac{\overline{\mathrm{mDSS}} - \mathsf{E}_0(\overline{\mathrm{mDSS}})}{\mathsf{Var}_0(\overline{\mathrm{mDSS}})^{1/2}} \underset{\mathrm{H}_0}{\overset{a}{\sim}} \mathcal{N}(0,1),$$

where

$$\mathsf{E}_0(\overline{\mathrm{mDSS}}) = m + \frac{1}{n}\sum_{i}^{n} \ln|\boldsymbol{\Sigma}_{P_i}| \quad \text{and} \quad \mathsf{Var}_0(\overline{\mathrm{mDSS}}) = \frac{2m}{n}.$$

3.1.3. Regression test

We can apply the regression approach to mDSS by setting $s_i = \mathrm{mDSS}_i$ in (3.5). We denote the test statistic $T_{\mathrm{mDSS}} \underset{\mathrm{H}_0}{\overset{a}{\sim}} \chi^2(2)$ as in (3.6).

3.2. Tests for iDSS

For n multivariate forecasts of dimension m, mn iDSSs can be computed:

$$\begin{pmatrix} \mathrm{iDSS}_{11} & \cdots & \mathrm{iDSS}_{1j} & \cdots & \mathrm{iDSS}_{1m} \\ \vdots & & \vdots & & \vdots \\ \mathrm{iDSS}_{i1} & \cdots & \mathrm{iDSS}_{ij} & \cdots & \mathrm{iDSS}_{im} \\ \vdots & & \vdots & & \vdots \\ \mathrm{iDSS}_{n1} & \cdots & \mathrm{iDSS}_{nj} & \cdots & \mathrm{iDSS}_{nm} \end{pmatrix}. \tag{3.8}$$

In contrast to mDSS, calibration tests for iDSS require multivariate testing techniques. We now propose three different approaches for multivariate testing which can be applied to the unconditional test and the regression test: In *Approach 1*, we test each component separately and adjust for multiple testing using Fisher's method. In *Approach 2*, we incorporate the correlation structure of the components into one test statistic. In *Approach 3*, we reduce the dimension from m to one composite variable and test the composite variable, for example the row average of (3.8) in m dimensions: $\overline{\text{iDSS}}_{i\cdot} = 1/m \sum_{j=1}^{m} \text{iDSS}_{ij}$.

3.2.1. Unconditional Tests

The idea of *Approach 1* is to first conduct the calibration tests (either unconditional test or regression test) on each component $j = 1, \ldots, m$, and then to adjust for multiplicity based on the p-values of the m tests. For each unconditional test, the statistic $z_{\text{iDSS}_{\cdot j}}$ is based on (3.3) by replacing $\overline{s}$ with the column mean of the matrix (3.8): $\overline{\text{iDSS}}_{\cdot j} = 1/n \sum_{i=1}^{n} \text{iDSS}_{ij}$ for each component j. We propose to use Fisher's method [17] to adjust for multiplicity, which combines the p-values p_j into a test statistic $R = -2 \sum_{j=1}^{m} \ln p_j \underset{\text{H}_0}{\sim} \chi^2(2m)$. A limitation of Fisher's method lies in the assumption of the independence of the tests, which is further discussed in Section 6.

To avoid the independence assumption of *Approach 1*, *Approach 2* takes the correlation between the iDSSs into account. Instead of separately testing each column average in $z_{\text{iDSS}_{\cdot j}}$, we construct a joint test by including all the $\overline{\text{iDSS}}_{\cdot j}$ and the correlations into a Hotelling-type statistic [24]:

$$Z_H^2 = n(\overline{\mathbf{iDSS}} - \mathsf{E}_0(\overline{\mathbf{iDSS}}))^\top (2\,\mathbf{\Omega})^{-1} (\overline{\mathbf{iDSS}} - \mathsf{E}_0(\overline{\mathbf{iDSS}})),$$

where

$$\begin{aligned} \overline{\mathbf{iDSS}} &= (\overline{\text{iDSS}}_{\cdot 1}, \ldots, \overline{\text{iDSS}}_{\cdot m})^\top \text{ and} \\ \mathsf{E}_0(\overline{\mathbf{iDSS}}) &= (\mathsf{E}_0(\overline{\text{iDSS}}_{\cdot 1}), \ldots, \mathsf{E}_0(\overline{\text{iDSS}}_{\cdot m}))^\top. \end{aligned}$$

For a sufficiently large number of observations n, we can thus assume $Z_H^2 \overset{a}{\underset{\mathrm{H}_0}{\sim}} \chi^2(m)$.

An alternative method for multivariate modelling is to reduce the number of random variables as in *Approach 3*. One common approach is to reduce the m-variate vector into a single quantity, here we use the row average $\overline{\mathrm{iDSS}}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{iDSS}_{ij}$. A test statistic can be formulated as in (3.4) by replacing s_i with $\overline{\mathrm{iDSS}}_{i\cdot}$,

$$z_{\mathrm{iDSS}_{i\cdot}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\overline{\mathrm{iDSS}}_{i\cdot} - \mathsf{E}_0(\overline{\mathrm{iDSS}}_{i\cdot})}{\mathsf{Var}_0(\overline{\mathrm{iDSS}}_{i\cdot})^{1/2}} \overset{a}{\underset{\mathrm{H}_0}{\sim}} \mathcal{N}(0,1),$$

where we have

$$\mathsf{E}_0(\overline{\mathrm{iDSS}}_{i\cdot}) = 1 + \frac{1}{m}\sum_{j=1}^{m} \ln \sigma_{ij}^2 \quad \text{and}$$

$$\mathsf{Var}_0(\overline{\mathrm{iDSS}}_{i\cdot}) = \frac{2}{m^2}\left(m + 2\sum_{j=1}^{m}\sum_{k>j}^{m} \rho_{jk}^2\right).$$

Note that the variance $\mathsf{Var}_0(\overline{\mathrm{iDSS}}_{i\cdot})$ is the same for all $i = 1, \ldots, n$.

3.2.2. Regression Tests

Approaches 1 - 3 can also be applied to the regression test in a similar way as the unconditional tests. In *Approach 1*, we regress the columns of the matrix (3.8) on their expectations separately, and get the statistic $T_{\cdot j} \overset{a}{\underset{\mathrm{H}_0}{\sim}} \chi^2(2)$ as in (3.6). We apply Fisher's method to the corresponding m p-values to obtain an overall p-value, following the same procedure as in the unconditional test.

In *Approach 2*, we use all entries iDSS_{ij} of (3.8) under incorporation of the correlation of the residual terms ϵ_{ij} in the regression (3.5). Specifically, we have $\mathsf{E}_0(\epsilon_{ij}) = 0$ and covariances

$$\begin{aligned} \mathrm{Cov}_0(\epsilon_{ij}, \epsilon_{lk}) &= 0, \quad \text{if } i \neq l, \\ \mathrm{Cov}_0(\epsilon_{ij}, \epsilon_{lk}) &\neq 0, \quad \text{if } i = l, \end{aligned}$$

so a generalised least squares estimate of the coefficients can be obtained and a Wald test statistic T_{iDSS} can be computed.

In *Approach 3*, we regress the composite score $\overline{\text{iDSS}}_{i\cdot}$ on its expectation $\mathsf{E}_0(\overline{\text{iDSS}}_{i\cdot})$ in (3.6) so the residuals are now independent. The Wald test statistic (3.6) based on this regression can be defined analogously.

4. Power evaluation

In practice, the data generating distribution remains hypothetical. The generator can be from the same distribution of forecasts or from a different distribution family, for example a t-distribution. We consider both situations in the power evaluation. In this section, we investigate the situation where the generator process and the forecasts are both from Gaussian distributions. In Section 4.2.2 a simulation study is presented, where the data generator follows a non-normal distribution.

When the generator is normally distributed, a miscalibrated forecast can be with an incorrect mean or incorrect covariance or both. Here in this section, we consider the two basic scenarios: forecasts with the correct covariance matrix and an incorrect mean; and forecasts with the correct mean and an incorrect covariance matrix. We would like to investigate the power of each test to reject the forecasts under these two scenarios: either the mean or the covariance is predicted incorrectly. Recall that the true and forecast means, as well as the corresponding covariance matrices are allowed to vary with the observations. For $i = 1, \ldots, n$, let $\boldsymbol{\Sigma}_i$ denote a diagonal matrix with entries σ_{ij}^2 for $1 \leq j \leq m$. Let $\boldsymbol{C}_\rho$ denote a uniform correlation matrix with correlation coefficient ρ, *i.e.*, for $1 \leq i, j \leq m$

$$(\boldsymbol{C}_\rho)_{ij} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{otherwise.} \end{cases} \tag{4.9}$$

In the following, we will use the fact that the eigenvalues of $\boldsymbol{C}_\rho$ are $\lambda_1 = 1 + (m-1)\rho$ and $\lambda_2 = \ldots = \lambda_m = 1 - \rho$ [29]. The scenarios considered in

the following are summarized in Table 1 below.

Scenario	Mean	Covariance
I	$\boldsymbol{\mu}_{P_i} \neq \boldsymbol{\mu}_{X_i}$	$\boldsymbol{\Sigma}_{P_i} = \boldsymbol{\Sigma}_{X_i}$
IIa	$\boldsymbol{\mu}_{P_i} = \boldsymbol{\mu}_{X_i}$	$\boldsymbol{\Sigma}_{P_i} = r^2 \boldsymbol{\Sigma}_{X_i}$ with $r \neq 1$
IIb	$\boldsymbol{\mu}_{P_i} = \boldsymbol{\mu}_{X_i}$	$\boldsymbol{\Sigma}_{X_i} = \boldsymbol{\Sigma}_i^{1/2} \boldsymbol{C}_\rho \boldsymbol{\Sigma}_i^{1/2}$ $\boldsymbol{\Sigma}_{P_i} = \boldsymbol{\Sigma}_i^{1/2} \boldsymbol{C}_\gamma \boldsymbol{\Sigma}_i^{1/2}$ with $\gamma \neq \rho$ and $\gamma \neq 0$
IIc	$\boldsymbol{\mu}_{P_i} = \boldsymbol{\mu}_{X_i}$	$\boldsymbol{\Sigma}_{X_i} = \boldsymbol{\Sigma}_i^{1/2} \boldsymbol{C}_\rho \boldsymbol{\Sigma}_i^{1/2}$ $\boldsymbol{\Sigma}_{P_i} = \boldsymbol{\Sigma}_i$ with $\rho \neq 0$

Table 1: Different scenarios to evaluate the power of the proposed calibration tests. Here $\boldsymbol{\Sigma}_i$ denotes a diagonal matrix whereas $\boldsymbol{C}_\rho$ and $\boldsymbol{C}_\gamma$ are both uniform correlation matrices.

Note that three different sub-cases of miscalibration are considered under scenario II: in scenario IIa the forecasts are wrong with respect to the covariances and the true covariance structure is allowed to vary with the observations. In scenario IIb and IIc, the forecasts are assumed to be wrong about the correlation structure but not the variances: in IIb, the forecast correlations are equal to $\gamma \neq 0$, whereas the true correlation ρ is different from γ and can be zero. The case $\rho \neq \gamma = 0$ is studied in scenario IIc. The analysis of the power based on explicit analytic formulas reveals that in this last case, where independence is wrongly assumed by the forecaster, the power converges to an asymptotic limit (for $n \to \infty$) which could be well below 1. The opposite situation, where the true data are in fact independent whereas the forecaster imposes a non-zero correlation (a special case of scenario IIb), is much more easily detected by the tests.

4.1. Test based on mDSS

Theorem 4.1. *Suppose that* $\boldsymbol{X}$ *follows a multivariate normal distribution* $\boldsymbol{\mathcal{N}}_m(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$*, while the predictive distribution* $\boldsymbol{P} = \boldsymbol{\mathcal{N}}_m(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ *is miscalibrated. Let* $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_X^{1/2}$ *such that* $\boldsymbol{\Sigma}_X = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$*,* $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ *be the diagonal matrix of the eigenvalues of* $\boldsymbol{\Gamma}^\top \boldsymbol{\Sigma}_P^{-1} \boldsymbol{\Gamma}$ *and* $\boldsymbol{L}$ *an orthogonal matrix such that* $\boldsymbol{\Gamma}^\top \boldsymbol{\Sigma}_P^{-1} \boldsymbol{\Gamma} = \boldsymbol{L}^\top \boldsymbol{\Lambda} \boldsymbol{L}$*. Then,*

$$\mathrm{mDSS}(\boldsymbol{X}, \boldsymbol{P}) \sim \sum_{j=1}^{m} \lambda_j Y_j$$

where $Y_1, \ldots, Y_m$ *are independent such that* $Y_i \sim \chi^2(1, b_j^2)$ *with*

$$(b_1, \ldots, b_m)^\top = \boldsymbol{L}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_P),$$

here $\chi^2(d, \tau)$ *denotes the non-central* χ^2*-distribution with* d *degrees of freedom and non-centrality parameter* τ*.*

Proof. See Appendix B. □

Theorem 4.1 allows us to derive the distribution of the test statistic (3.7), and to calculate explicitly the asymptotic power (for $n \to \infty$) under many alternative hypotheses. Given that the i-th forecast is $\mathcal{N}(\boldsymbol{\mu}_{P_i}, \boldsymbol{\Sigma}_{P_i})$, the null hypothesis is

$$\mathrm{H}_0 : \boldsymbol{\mu}_{P_i} = \boldsymbol{\mu}_{X_i} \quad \text{and} \quad \boldsymbol{\Sigma}_{P_i} = \boldsymbol{\Sigma}_{X_i}, \quad \text{for } i = 1, \ldots, n.$$

Under H_0, $Z_{\mathrm{mDSS}} \sim \chi^2(mn)$. Under any fixed alternative H_1, we know that Z_{mDSS} is distributed as a weighted sum of mn independent non-central chi-squared random variables. It follows from Theorem 4.1 that the weights are given by the eigenvalues of $\boldsymbol{\Sigma}_{X_i}^{1/2} \boldsymbol{\Sigma}_{P_i}^{-1} \boldsymbol{\Sigma}_{X_i}^{1/2}$ whereas the non-centrality parameters, $\nu_{ij}, 1 \le i \le n, 1 \le j \le m$ involve $\boldsymbol{\Sigma}_{X_i}^{1/2}, \boldsymbol{\Sigma}_{P_i}^{-1}$ and $\boldsymbol{\mu}_{P_i} - \boldsymbol{\mu}_{X_i}$. In particular, it follows from the proof of Theorem 4.1 that

$$\sum_{j=1}^{m} \nu_{ij} = (\boldsymbol{\mu}_{P_i} - \boldsymbol{\mu}_{X_i})^\top \boldsymbol{\Sigma}_{X_i}^{-1} (\boldsymbol{\mu}_{P_i} - \boldsymbol{\mu}_{X_i}). \tag{4.10}$$

For small n, the power can be computed as

$$\begin{aligned} & \Pr\left(Z_{\text{mDSS}} > \chi^2_{1-\alpha/2}(mn) \,|\, \text{H}_1\right) + \Pr\left(Z_{\text{mDSS}} < \chi^2_{\alpha/2}(mn) \,|\, \text{H}_1\right) \\ = \; & \Pr\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_{ij}\ \chi^2(1,\nu_{ij}) \ > \chi^2_{1-\alpha/2}(mn)\right) \\ + \; & \Pr\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_{ij}\ \chi^2(1,\nu_{ij}) < \chi^2_{\alpha/2}(mn)\right) \end{aligned}$$

where $\chi^2_\gamma(mn)$ is the γ-quantile of the $\chi^2(mn)$-distribution. Practically, we compute the probability above via numerical methods using the R-package `survey` [28] for any specified precision.

For the unconditional test based on z_{mDSS}, the asymptotic power is given by

$$\Pr\left(|z_{\text{mDSS}}| > z_{1-\alpha/2} \,|\, \text{H}_1\right),$$

where z_γ is the γ-quantile of $\mathcal{N}(0,1)$. In the following, we give the distribution under the alternative hypothesis in each of the scenarios of Table 1 and also the asymptotic power as $n \to \infty$. To do so, we define

$$W_i \;:=\; (\boldsymbol{X}_i - \boldsymbol{\mu}_{P_i})^\top \boldsymbol{\Sigma}_{P_i}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_{P_i})$$

and let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \boldsymbol{I}_m)$ denote independent multivariate standard normal random vectors.

Scenario I. Let $\Delta\boldsymbol{\mu}_i = \boldsymbol{\mu}_{P_i} - \boldsymbol{\mu}_{X_i} \neq \boldsymbol{0}$. We then have

$$\begin{aligned} W_i \;&\overset{d}{=}\; \{\boldsymbol{Y}_i - \boldsymbol{\Sigma}_{X_i}^{-1/2}(\Delta\boldsymbol{\mu}_i)\}^\top \{\boldsymbol{Y}_i - \boldsymbol{\Sigma}_{X_i}^{-1/2}(\Delta\boldsymbol{\mu}_i)\} \\ &\sim\; \chi^2\left(m, (\Delta\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_{X_i}^{-1}(\Delta\boldsymbol{\mu}_i)\right). \end{aligned}$$

The obtained distribution is aligned with the expression given in (4.10) since the weights $\lambda_{ij} = 1, 1 \leq j \leq m$, imply that the non-centrality parameters are all equal to the scalar product $(\Delta\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_{X_i}^{-1}(\Delta\boldsymbol{\mu}_i)$. It follows that

$$Z_{\text{mDSS}} \sim \chi^2\left(mn, \sum_{i=1}^{n} (\Delta\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_{X_i}^{-1}(\Delta\boldsymbol{\mu}_i)\right).$$

In the following, we assume that $\Delta\boldsymbol{\mu}_i = (\Delta\mu, \ldots, \Delta\mu)^\top$ with $\Delta\mu \in \mathbb{R}$ and independent of i and that $\boldsymbol{\Sigma}_{X_i} = \boldsymbol{C}_\rho$ as defined in (4.9). Then it is not difficult to show that the entries of $\boldsymbol{R} = \boldsymbol{C}_\rho^{-1}$ are given by

$$R_{ij} = \begin{cases} \theta & \text{if } i = j \\ \theta\kappa & \text{otherwise} \end{cases}$$

where

$$\theta = \frac{1 + (m-2)\rho}{(1-\rho)(1+(m-1)\rho)} \quad \text{and} \quad \kappa = -\frac{\rho}{1+(m-2)\rho}.$$

It follows that $(\Delta\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_{X_i}^{-1}(\Delta\boldsymbol{\mu}_i) = m(\Delta\mu)^2\theta\{1+(m-1)\kappa\}$, so

$$\begin{aligned} W_i &\sim \chi^2\left(m, \frac{m}{1+(m-1)\rho}(\Delta\mu)^2\right) \text{ and} \\ Z_{\text{mDSS}} &\sim \chi^2\left(mn, \frac{mn}{1+(m-1)\rho}(\Delta\mu)^2\right). \end{aligned}$$

Therefore, for $i = 1, \ldots, n$,

$$\begin{aligned} \mathsf{E}(W_i) &= m\left\{1 + \frac{(\Delta\mu)^2}{1+(m-1)\rho}\right\} \neq m, \text{ and} \\ \mathsf{Var}(W_i) &= 2m\left\{1 + 2\,\frac{(\Delta\mu)^2}{1+(m-1)\rho}\right\} < \infty. \end{aligned}$$

It follows from the Proposition in Appendix C that the asymptotic power is

$$\lim_{n\to\infty} \Pr\left(\frac{\sqrt{n}\,|Z_{\text{mDSS}}/n - m|}{\sqrt{2m}} > z_{1-\alpha/2}\right) = 1.$$

Scenario IIa. Now we have

$$W_i \sim \frac{1}{r^2}\chi^2(m), \text{ and } Z_{\text{mDSS}} \sim \frac{1}{r^2}\chi^2(mn).$$

Hence the weights are $\lambda_j = 1/r^2$ for $j = 1, \ldots, m$ and the non-centrality parameters are all equal to 0. Therefore, for $i = 1, \ldots, n$,

$$\mathsf{E}(W_i) = m/r^2 \neq m, \text{ and } \mathsf{Var}(W_i) = 2m/r^4 < \infty.$$

It follows from the Proposition in Appendix C that the asymptotic power is

$$\lim_{n\to\infty} \Pr\left(\frac{\sqrt{n}\,|Z_{\mathrm{mDSS}}/n - m|}{\sqrt{2m}} > z_{1-\alpha/2}\right) = 1.$$

To study scenario IIb and IIc, we will develop our calculations in the general case $\boldsymbol{\Sigma}_{X_i} = \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{C}_\rho\boldsymbol{\Sigma}_i^{1/2}$ and $\boldsymbol{\Sigma}_{P_i} = \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{C}_\gamma\boldsymbol{\Sigma}_i^{1/2}$. Recall that $\boldsymbol{\Sigma}_i$ is a diagonal matrix with positive diagonal elements σ_{ij}^2, $1 \le j \le m$. Now, we can easily check that $\boldsymbol{C}_\rho\boldsymbol{C}_\gamma = \boldsymbol{C}_\gamma\boldsymbol{C}_\rho$ for any correlation coefficients ρ and γ. Hence, there exists an orthogonal matrix $\boldsymbol{Q}$ such that

$$\boldsymbol{C}_\rho = \boldsymbol{Q}\boldsymbol{\Delta}_\rho\boldsymbol{Q}^\top, \quad \text{and} \quad \boldsymbol{C}_\gamma = \boldsymbol{Q}\boldsymbol{\Delta}_\gamma\boldsymbol{Q}^\top$$

where, for $\theta \in (-1,1)$, $\boldsymbol{\Delta}_\theta$ is the diagonal $m \times m$ matrix with diagonal elements $\boldsymbol{\Delta}_{\theta,11} = 1+(m-1)\theta$ and $\boldsymbol{\Delta}_{\theta,ii} = 1-\theta$ for $2 \le i \le m$. Also, let $\boldsymbol{\Gamma}_i = \boldsymbol{\Sigma}_{X_i}^{1/2} = \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{Q}\boldsymbol{\Delta}_\rho^{1/2}$ so that $\boldsymbol{\Sigma}_{X_i} = \boldsymbol{\Gamma}_i\boldsymbol{\Gamma}_i^\top$. Then, with $\boldsymbol{Y}_i = (Y_{i1},\ldots,Y_{im})^\top = \boldsymbol{\Gamma}_i^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_{X_i}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$ and $\boldsymbol{\mu}_{P_i} = \boldsymbol{\mu}_{X_i}$, we have

$$\begin{aligned}
W_i &= \{\boldsymbol{\Gamma}_i^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_{X_i})\}^\top\boldsymbol{\Gamma}_i^\top\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{Q}\boldsymbol{\Delta}_\gamma^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\Gamma}_i\{\boldsymbol{\Gamma}_i^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_{X_i})\} \\
&= \boldsymbol{Y}_i^\top\boldsymbol{\Gamma}_i^\top\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{Q}\boldsymbol{\Delta}_\gamma^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\Gamma}_i\boldsymbol{Y}_i \\
&= \boldsymbol{Y}_i^\top\boldsymbol{\Delta}_\rho^{1/2}\boldsymbol{Q}^\top\boldsymbol{\Sigma}_i^{1/2}\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{Q}\boldsymbol{\Delta}_\gamma^{-1}\boldsymbol{Q}^\top\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\Sigma}_i^{1/2}\boldsymbol{Q}\boldsymbol{\Delta}_\rho^{1/2}\boldsymbol{Y}_i \\
&= \boldsymbol{Y}_i^\top\boldsymbol{\Delta}_\rho^{1/2}\boldsymbol{\Delta}_\gamma^{-1}\boldsymbol{\Delta}_\rho^{1/2}\boldsymbol{Y}_i \\
&= \boldsymbol{Y}_i^\top\boldsymbol{\Delta}_\rho\boldsymbol{\Delta}_\gamma^{-1}\boldsymbol{Y}_i \\
&= \frac{1+(m-1)\rho}{1+(m-1)\gamma}\,Y_{i1}^2 + \frac{1-\rho}{1-\gamma}\sum_{j=2}^{m} Y_{ij}^2 \\
&\sim \frac{1+(m-1)\rho}{1+(m-1)\gamma}\,\chi^2(1) + \frac{1-\rho}{1-\gamma}\,\chi^2(m-1)
\end{aligned}$$

and

$$Z_{\mathrm{mDSS}} \sim \frac{1+(m-1)\rho}{1+(m-1)\gamma}\,\chi^2(n) + \frac{1-\rho}{1-\gamma}\,\chi^2((m-1)n).$$

Scenario IIb. Here we assume $\gamma \neq 0$ and $\rho \neq \gamma$. Then, according to the previous calculations we have, for $i = 1, \ldots, n$,

$$\begin{aligned} \mathsf{E}(W_i) &= \frac{1+(m-1)\rho}{1+(m-1)\gamma} + (m-1)\frac{1-\rho}{1-\gamma} \neq m, \text{ and} \\ \mathsf{Var}(W_i) &= 2\left[\frac{\{1+(m-1)\rho\}^2}{\{1+(m-1)\gamma\}^2} + \frac{(1-\rho)^2}{(1-\gamma)^2}\right] < \infty. \end{aligned}$$

It follows from the Proposition in Appendix C that the asymptotic power is

$$\lim_{n\to\infty} \Pr\left(\frac{\sqrt{n}\ |Z_{\mathrm{mDSS}}/n - m|}{\sqrt{2m}} > z_{1-\alpha/2}\right) = 1.$$

Scenario IIc. In this scenario $\gamma = 0$ and hence

$$\begin{aligned} W_i &\sim \{1+(m-1)\rho\}\chi^2(1) + (1-\rho)\chi^2(m-1) \text{ and} \\ Z_{\mathrm{mDSS}} &\sim \{1+(m-1)\rho\}\,\chi^2(n) + (1-\rho)\,\chi^2((m-1)n). \end{aligned}$$

For $i = 1, \ldots, n$, we have $\mathsf{E}(W_i) = m$ and

$$\mathsf{Var}(W_i) = 2\{(1+(m-1)\rho)^2 + (m-1)(1-\rho)^2\} = \sigma^2 < \infty,$$

so $\sigma^2/(2m) = 1 + (m-1)\rho^2$. It follows from the Proposition in Appendix C that the asymptotic power is

$$\begin{aligned} &\lim_{n\to\infty} \Pr\left(\frac{\sqrt{n}\ |Z_{\mathrm{mDSS}}/n - m|}{\sqrt{2m}} > z_{1-\alpha/2}\right) \\ &= \Pr\left(|Z|\frac{\sigma}{\sqrt{2m}} > z_{1-\alpha/2}\right) \\ &= 2\left\{1 - \Phi\left(\frac{z_{1-\alpha/2}}{\sqrt{1+(m-1)\rho^2}}\right)\right\}, \end{aligned} \tag{4.11}$$

where Φ denotes the CDF of $Z \sim \mathcal{N}(0,1)$.

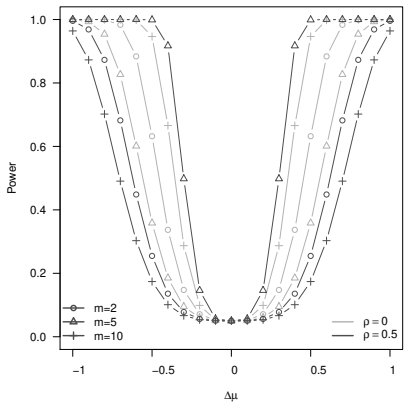
In Figure 1, we plot the power of the test based on Z_{mDSS} for scenarios I-IIc and different dimensions. Here, we take $m = 2, 5$ and 10. In each scenario, the power is presented as a function of the parameter determining

the distance between H_0 and H_1 for the sample size $n = 100$. The power function for the scenarios I, IIa and IIb takes a U-shape as the difference between H_0 and H_1 increases. The difference between H_0 and H_1 corresponds to $|\triangle\mu|$ in Figure 1a, σ_P/σ_X in Figure 1b and ρ in Figure 1c and 1d. The hypotheses H_0 and H_1 are indistinguishable when $\triangle\mu = 0$, $\sigma_P/\sigma_X = 1$ or $\rho = 0$, and the power is equal to the significance level $\alpha = 0.05$ as expected. Note that in Figure 1b, the power does not involve the correlation coefficient ρ and hence only the three lines corresponding to the dimensions $m = 2$, $m = 5$ and $m = 10$ are plotted. In Figures 1c and 1d, corresponding to the scenarios IIb and IIc, ρ is taken to be bigger than $-1/4$ and $-1/9$ respectively so that the covariance matrix is positive definite for $m = 5$ and $m = 10$. As opposed to scenarios I, IIa and IIb for which the power increases quickly, the power remains rather low in the case of scenario IIc. This is expected by our analysis above where we show that, in this case, the power should converge to (4.11). Taking for example $m = 10$, this result means that even in the extreme case where $\rho = \pm 1$, having a forecast that assumes independence between the observations yields a test with an asymptotic power that is not larger than 53.54%. Note however that the power increases to 1 as the dimension m increases. For example, if $\rho = \pm 0.5$ and $m = 200$, the asymptotic power is 78.32%.

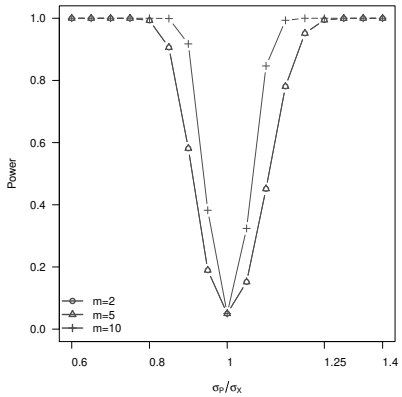
We note that the unconditional test based on z_{mDSS} is almost as powerful as the one based on Z_{mDSS} when n is sufficiently large. We have investigated the asymptotic power based on $n = 100$, and found results that are similar to those shown in Figure 1 for scenarios I-IIc. Additional details are given in the Appendix.

4.2. Simulation studies

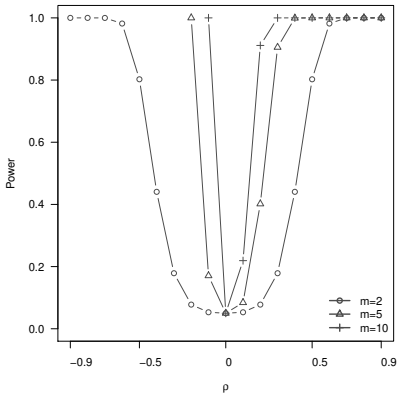
The power of each test based on iDSS is evaluated via simulations. To assess this power for miscalibrated forecasts, we have simulated 10 000 forecasts for $n \in \{100, 500\}$ and the dimensions $m = 2, 5, 10$. The power of each test is estimated as the proportion of rejected null hypotheses based on a



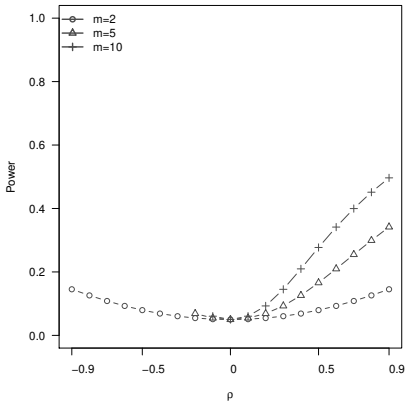
(a) Scenario I with $\sigma_P = \sigma = 1$



(b) Scenario IIa



(c) Scenario IIb



(d) Scenario IIc

Figure 1: Power of the mDSS test based on χ^2-distribution in scenario I, IIa, IIb and IIc for $m = 2, 5, 10$.

significance level of $\alpha = 5\%$. Note that with the chosen number of simulated forecasts, the Monte Carlo standard error of the empirical power estimates is always smaller than 0.01.

4.2.1. Gaussian data generating distribution

Suppose the true data generating distribution is $\boldsymbol{X}_i \sim \boldsymbol{\mathcal{N}}_m(\boldsymbol{\mu}_{X_i}, \boldsymbol{\Sigma}_{X_i})$, where the means $\boldsymbol{\mu}_{X_i}$ are realisations of m independent standard normal random variables; the components of the variance vector $\boldsymbol{\sigma}^2_{X_i}$ are independently generated from the $\chi^2(4)$ distribution and ρ_{X_i} is taken to be either equal to 0 or 0.5 for $i = 1, \ldots, n$. For scenario I, forecasts are made with incorrect mean $\boldsymbol{\mu}_{P_i} = \boldsymbol{\mu}_{X_i} \pm 0.2\boldsymbol{\sigma}_{X_i}$ and $\boldsymbol{\Sigma}_{P_i} = \boldsymbol{\Sigma}_{X_i}$. For scenarios IIa - IIc, with incorrectly predicted covariance matrix, we have fixed $\boldsymbol{\mu}_{P_i} = \boldsymbol{\mu}_{X_i}$ but $\boldsymbol{\Sigma}_{P_i} \neq \boldsymbol{\Sigma}_{X_i}$. To be specific, for scenario IIa, we have set $\boldsymbol{\sigma}^2_{P_i} = 1.44\boldsymbol{\sigma}^2_{X_i}$ and $\rho_{P_i} = \rho_{X_i}$; for scenario IIb and IIc, $\boldsymbol{\sigma}^2_{P_i} = \boldsymbol{\sigma}^2_{X_i}$ and $\rho_{P_i} \neq \rho_{X_i}$, where ρ_{P_i} and ρ_{X_i} take either the value 0 or 0.5. The simulation was carried out for $m = 2, 5$ and 10.

Figures 2 to 4 display the proportions of rejections of the null hypotheses of each test in scenarios I, IIa, IIb and IIc. Overall the power of all the tests increases as the number of observations n or the dimension m increases, except in Figure 4b. For scenario I and IIa, the power increases quickly, such that almost all the tests have a power greater than 80% for scenario I and equal to 100% for scenario IIa when $n = 500$. The regressions tests are more powerful than unconditional tests, which has also been observed in Held et al. [23]. Note that in Figure 2a and 3a with $\rho = 0$, the tests based on iDSS in *Approach 3* are equivalent to the tests based on mDSS, since mDSS and iDSS are coherent. The statistics based on mDSS and the row average of iDSS are always equal, which holds for both the unconditional test and the regression test.

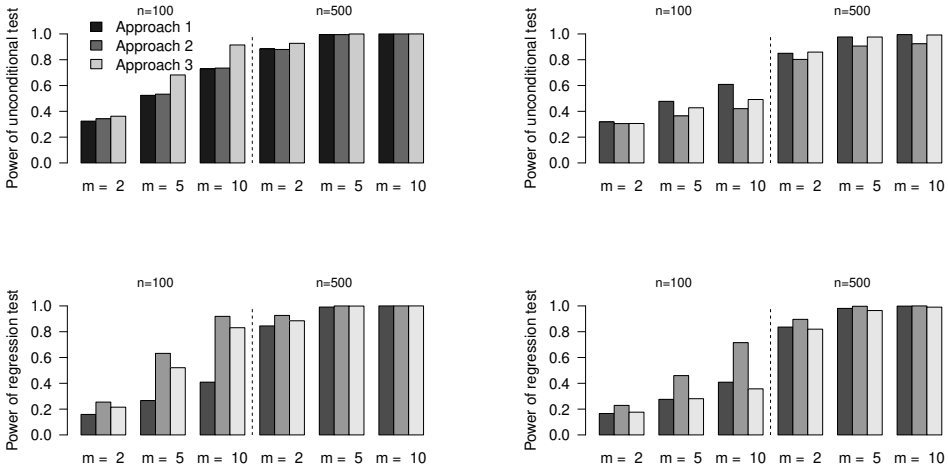
For scenarios IIb and IIc, where the forecasts are marginally calibrated (the marginal forecast distribution of each component is equal to the marginal true data generating distribution) whereas the correlations are wrongly pre-

	$\rho = 0$		$\rho = 0.5$	
	$n = 100$	$n = 500$	$n = 100$	$n = 500$
z_{mDSS}	-3.055	-6.837	-3.055	-6.837
$z_{\text{iDSS}_{\cdot 1}}$	-2.160	-4.833	-2.150	-4.833
$z_{\text{iDSS}_{\cdot 2}}$	-2.160	-4.836	-2.166	-4.845
$z_{\text{iDSS}_{i\cdot}}$	-3.055	-6.837	-2.729	-6.121

Table 2: The average of each z-statistic from 10 000 rounds of simulation in scenario IIa: $\boldsymbol{\sigma}_P^2 = 1.44\boldsymbol{\sigma}_X^2$ for $m = 2$.

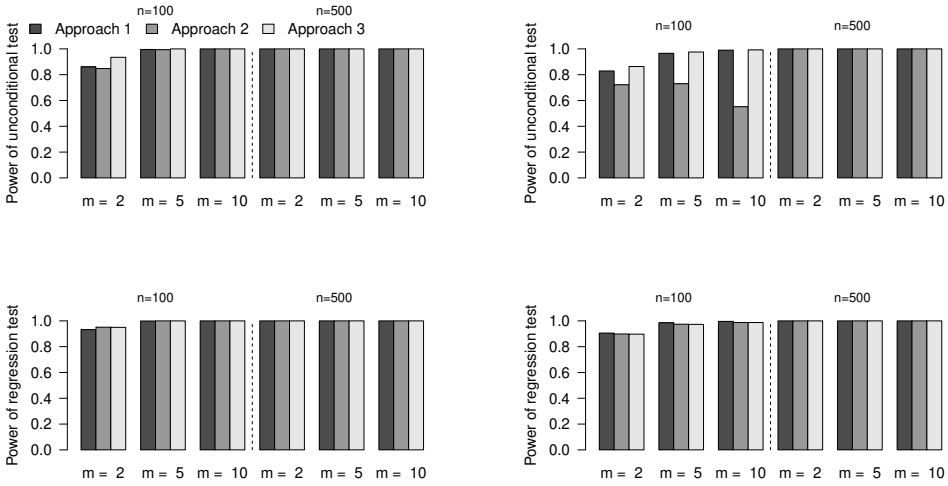
dicted, Figure 4 shows a lack of power for both the unconditional and regression tests. This result is reminiscent of the one found for the test based on mDSS in the case of scenario IIc (see Section 4.1), and it is interesting to note that this lack of power is observed also in the case where $\rho_X = 0$ and $\rho_P \neq 0$. For the test based on mDSS whose power was studied analytically, this case corresponds to scenario IIb where the power is shown to converge to 1 as $n \to \infty$.

As discussed in Section 3, the sign of the z-statistic provides hints as to whether the forecasts are overdispersed, underdispersed or biased in univariate forecasting. It is more complicated in multivariate forecasting, because in addition to the perfect marginal forecast for each component, the correlation should also be predicted correctly for an ideal multivariate forecast. The value of the z-statistic is positive if the forecasts are biased, underdispersed or strongly correlated. On the other hand, the z-value is negative if forecasts have larger variance or smaller correlation coefficients. For example, in the simulation in scenario IIa where $\boldsymbol{\sigma}_P^2 = 1.44\boldsymbol{\sigma}_X^2$ the average z-statistics are all negative (see Table 2). Further discussion is provided in Section 5.



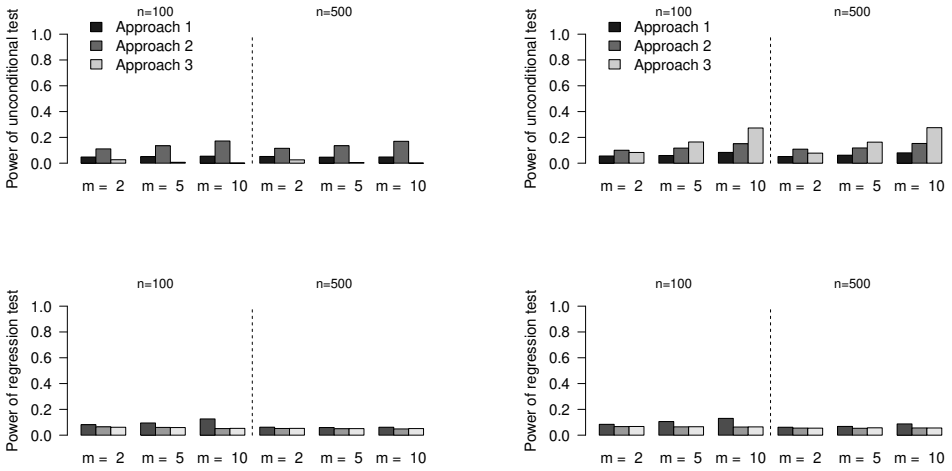
(a) $\rho = 0$ (b) $\rho = 0.5$

Figure 2: Power of calibration tests for miscalibration with different means (Scenario I).



(a) $\rho = 0$ (b) $\rho = 0.5$

Figure 3: Power of calibration tests for miscalibration with different scales $\sigma_P^2 = 1.44\sigma_X^2$ (Scenario IIa).



(a) $\rho_P = 0.5$ and $\rho_X = 0$ (Scenario IIb) (b) $\rho_P = 0$ and $\rho_X = 0.5$ (Scenario IIc)

Figure 4: Power of calibration tests for miscalibration with different correlation coefficients.

4.2.2. Multivariate t data generating distribution

In this simulation, we have used miscalibrated Gaussian forecasts with correct means (as zero) and incorrect covariances. The true data generator was chosen as a multivariate t-distribution with 3 degrees of freedom, mean vector zero and covariance matrix equal to $3\boldsymbol{C}_\rho$, where $\rho = 0$ or 0.5. The forecasts follow a $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$ distribution where we use $\boldsymbol{\Sigma}_i = \sigma_i^2 \boldsymbol{C}_\rho$ (for $i = 1, \ldots, n$), so the correlation ρ is predicted correctly. Here, σ_i^2 was sampled from a mixture of uniform distributions $\mathcal{U}(1, 2)$ and $\mathcal{U}(3, 4)$ with probability 1/2. Note that when $\sigma_i^2 \sim \mathcal{U}(1, 2)$ the forecast is underdispersed, and overdispersed in the other case.

Table 3 illustrates the power for each test. Overall, the results are very similar to Scenario IIa in Section 4.2. The power of every test increases as the dimension m increases and the number of observations n increases. All the tests reach more than 80% power when $n = 500$ for $m = 2$. In the regression test, the iDSS tests from Approach 2 and 3 yield the same power.

	n = 100			n = 500		
	m = 2	m = 5	m = 10	m = 2	m = 5	m = 10
Unconditional						
mDSS-1	59.8/59.8	71.1/71.1	79.4/79.4	83.2/83.2	92.7/92.7	95.2/95.2
mDSS-2	53.9/53.9	66.4/66.4	75.6/75.6	80.4/80.4	90.9/90.9	94.4/94.4
iDSS-App.1	60.1/58.5	80.2/75.4	92.2/87.2	84.1/82.3	95.9/93.1	98.9/97.7
iDSS-App.2	61.1/59.5	80.2/75.0	92.5/86.3	84.5/82.8	96.0/93.6	99.0/98.1
iDSS-App.3	53.9/51.2	66.4/56.3	75.6/59.1	80.4/77.8	90.9/83.5	94.4/86.1
Regression						
mDSS	50.5/50.5	61.0/61.0	66.0/66.0	86.7/86.7	90.9/90.9	91.6/91.6
iDSS-App.1	57.8/56.0	79.9/76.5	89.5/85.2	92.8/91.5	98.6/97.7	99.6/98.7
iDSS-App.2	50.5/47.7	61.0/54.9	66.0/58.1	86.7/85.2	90.9/89.2	91.6/89.8
iDSS-App.3	50.5/47.7	61.0/54.9	66.0/58.1	86.7/85.2	90.9/89.2	91.6/89.8

Table 3: Power (in %) of each test in the simulation when the true generator follows a multivariate t-distribution with 3 degrees of freedom. The two numbers separated by "/" are the power when $\rho = 0$ and $\rho = 0.5$. mDSS-1 is the test of mDSS based on χ^2-distribution, mDSS-2 is the unconditional test of mDSS, App.1–3 refer to tests of Approach 1–3 based on iDSS.

These two regression tests are equivalent, since the correlation structure of iDSS components are included in the two regressions in both approaches.

5. Evaluation of predictive models for weekly number of reported *Norovirus* infections in Germany

Norovirus is the most common cause of viral gastroenteritis in humans. The annual reported incidence of *Norovirus* infections was 142 cases per 100 000 inhabitants in 2011 in Germany. A lot of research has been done to understand the disease dynamics and further to initiate a surveillance system for early detection of future outbreaks [18; 27; 41]. Calibration tests

discussed in this paper can help to select good models for the predictions of future disease incidence. We extracted data on laboratory-confirmed weekly number of *Norovirus* infections reported between 2011 and 2014 from the national surveillance database in Germany (date of query: 17/02/2015). The data are freely available via `https://survstat.rki.de/`. Case numbers are given by week and gender, resulting in $2 \cdot 52 \cdot 4 = 416$ observations to be included in the analysis. Figure 5 shows the number of reported infections among males and females. Strong seasonal patterns can be observed in both time series: peaks during winter and troughs from spring to summer. Cases reported in calendar week 53 were randomly distributed to calendar week 52 of the same year or to calendar week 1 of the following year, respectively [5]. A sudden decrease in calendar week 52 or calendar week 1 occurs every year because there are not exactly 7 days in these two calendar weeks each year.

To understand the disease epidemic and further construct a model for outbreak predictions, we have applied the bivariate two-component model [34], assuming the reported incidence counts follow a negative binomial distribution. Compared with other time series methods, for example the vector autoregression method (VAR), this model allows for time-varying variances depending on the current mean. The two-component model is implemented in the R package `surveillance` [31]. Let $X_{1,t}$ and $X_{2,t}$ denote the number of cases among males and females, respectively, in week t. Our joint time series model is

$$X_{jt} \,|\, x_{1,t-1}, x_{2,t-1} \sim \mathcal{NB}(\mu_{jt}, \psi_j),$$

where $j = 1, 2$ represents gender, $t = 2, \ldots, 208$ represents week and

$$\begin{pmatrix} \mu_{1t} \\ \mu_{2t} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \phi \\ \phi & \lambda_2 \end{pmatrix} \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} + \begin{pmatrix} e_1\nu_{1t} \\ e_2\nu_{2t} \end{pmatrix}.$$

When the overdispersion parameter ψ_j goes to infinity, the negative binomial distribution converges to a Poisson distribution. The gender proportions e_j of males and females in the whole population [8] are included as offsets. The

endemic component $e_j\nu_{jt}$ describes the background risk of new events caused by external factors. The autoregressive component is driven by the observed cases $x_{1,t-1}$ and $x_{2,t-1}$ in the past week [31]. In our model, the endemic parameters ν_{1t} and ν_{2t} share the same seasonal pattern:

$$\nu_{jt} = \alpha + \beta\, \mathbf{1}_{(\text{calendar week 1 or 52})} + \delta \sin(\omega t) + \gamma \cos(\omega t),$$

where $\omega = 2\pi/52$. The artificial decrease in reporting incidence in calendar week 52 and calendar week 1 of each year is incorporated through inclusion of the indicator $\mathbf{1}_{(\text{calendar week 1 or 52})}$.

We tried different models with or without the autoregressive components λ_j or the interactive effect ϕ:

$$\begin{aligned}
&\text{Model A} : \lambda_1 = \lambda_2 = 0 \text{ and } \phi = 0,\\
&\text{Model B} : \lambda_1 = \lambda_2 = \lambda \text{ and } \phi = 0,\\
&\text{Model C} : \lambda_1, \lambda_2 > 0 \text{ and } \phi > 0.
\end{aligned}$$

Parameter estimates are given in Table 4 while Figure 6 illustrates the fit of the three models to the full data. Model A only includes the seasonal pattern without any autoregressive component and shows poor fitting results and strong overdispersion ($\hat{\psi}$ is small in Table 4). Model B includes the autoregressive effect within the same time series while Model C further allows for gender specific autoregressive effect and an interaction ϕ between the two time series. In comparison with Model A, the inclusion of the dependency term (λ or ϕ) of infections from the previous week largely improves the performance of the Models B and C.

The data from the first two years are used as a learning set and the data from the last two years are a validation set to compute one-step-ahead predictions [33]. In total 104 one-step-ahead predictions were computed, for computational simplicity based on the parameter estimates from fits to the full data (as shown in Table 4). The two-component model assumes that the two time series X_{1t} and X_{2t} are conditionally independent given

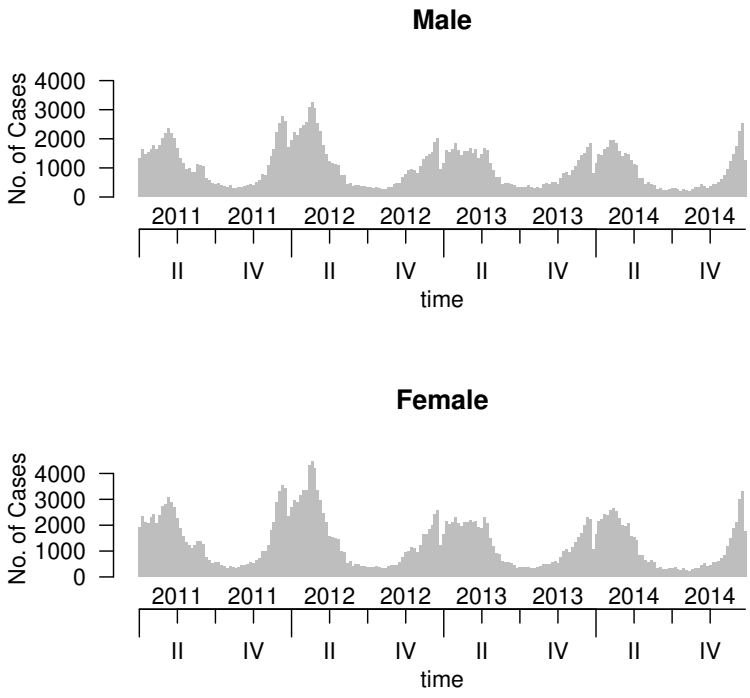


Figure 5: Number of reported infections among males and females in 2011-2014 in Germany

the past counts. However, the Pearson residuals, defined as $(x_{jt} - \mu_{jt})/\sigma_{jt}$ where $\sigma_{jt}^2 = \mu_{jt} + \mu_{jt}^2/\psi_j$, turn out to be highly correlated between males and females for all three models: $\hat{\rho} = 0.96$, 0.89, 0.88 for model A, B and C, respectively. The correlation coefficient between the two time series gives evidence that the two time series are not independent and so the dependency should be included in the predictive distribution. Therefore, we incorporate the correlation into the predictions and consider the estimated correlation $\hat{\rho}$ from the residuals in the learning set to be the correlation between the two outcomes (males and females). The joint prediction $\boldsymbol{P}$ is then based on the marginal distribution P_1 and P_2 and the correlation coefficient $\hat{\rho}$ estimated from the learning set.

Since there are no tools available yet to evaluate multivariate negative binomial predictions, we apply the calibration tests for multivariate Gaussian forecasts to the predictions for model A, B and C. For these data with large counts, a normal distribution having the same mean and variance as the negative binomial gives a good approximation to the CDF [2]. Table 5 gives p-values and z-values from the calibration tests discussed in this paper. The p-values of univariate calibration tests [23] are also shown in Table 5 for males and females separately. The p-value in *Approach 1* (denoted as iDSS-App.1 in Table 5) is computed from the two p-values of univariate tests using Fisher's method. Both the univariate tests and multivariate tests show that Model A is not well-calibrated. Further, the z-values of the univariate tests for "Male" and "Female" are both negative, indicating the forecasts are overdispersed, especially for females. In contrast, the z-value in the mDSS test equals 20.05. These differences of the z-values indicate that the correlation of the predictions ($\hat{\rho} = 0.96$) is too large. An interesting feature of the results shown in Table 5 is that the p-values from the tests based on mDSS and on iDSS are different for Model B. This finding is consistent with the result of the power evaluation in Section 4, which indicates that tests based on iDSS or univariate tests (see Table 5) have only little or no power to reject the

multivariate forecasts with incorrect predictive correlation. When the model is further improved to Model C, there is no evidence for miscalibration (using the conventional threshold of 0.05).

The result of the model evaluation give evidence that univariate calibration tests are not suitable for multivariate predictions. The corresponding scatterplots of mDSS with the fitted regression line are shown in Figure 7. Visual inspection indicates that model A is strongly miscalibrated, with a tendency for increasing score values with increasing determinant of the covariance matrix, *i.e.*, E_0(mDSS). For model B, the two regression lines are roughly parallel with a slight difference in the intercept, while model C seems to be sufficiently well calibrated.

Parameter	Model A	Model B	Model C
α	7.557 (0.016)	5.987 (0.126)	6.091 (0.120)
β	-0.422 (0.079)	-0.877 (0.221)	-0.848 (0.199)
δ	0.480 (0.020)	0.092 (0.067)	0.128 (0.061)
γ	0.930 (0.021)	1.194 (0.062)	1.158 (0.055)
λ_1		0.776 (0.027)	0.627 (0.110)
λ_2			0.713 (0.070)
ϕ			0.080 (0.087)
ψ_1	12.247 (1.315)	42.766 (4.503)	43.658 (4.568)
ψ_2	13.561 (1.455)	43.567 (4.540)	44.877 (4.645)

Table 4: Parameter estimates (with standard errors in brackets) from Model A–C.

6. Discussion

In this paper, we have proposed several significance tests to assess calibration of multivariate Gaussian predictions. In particular, we extended

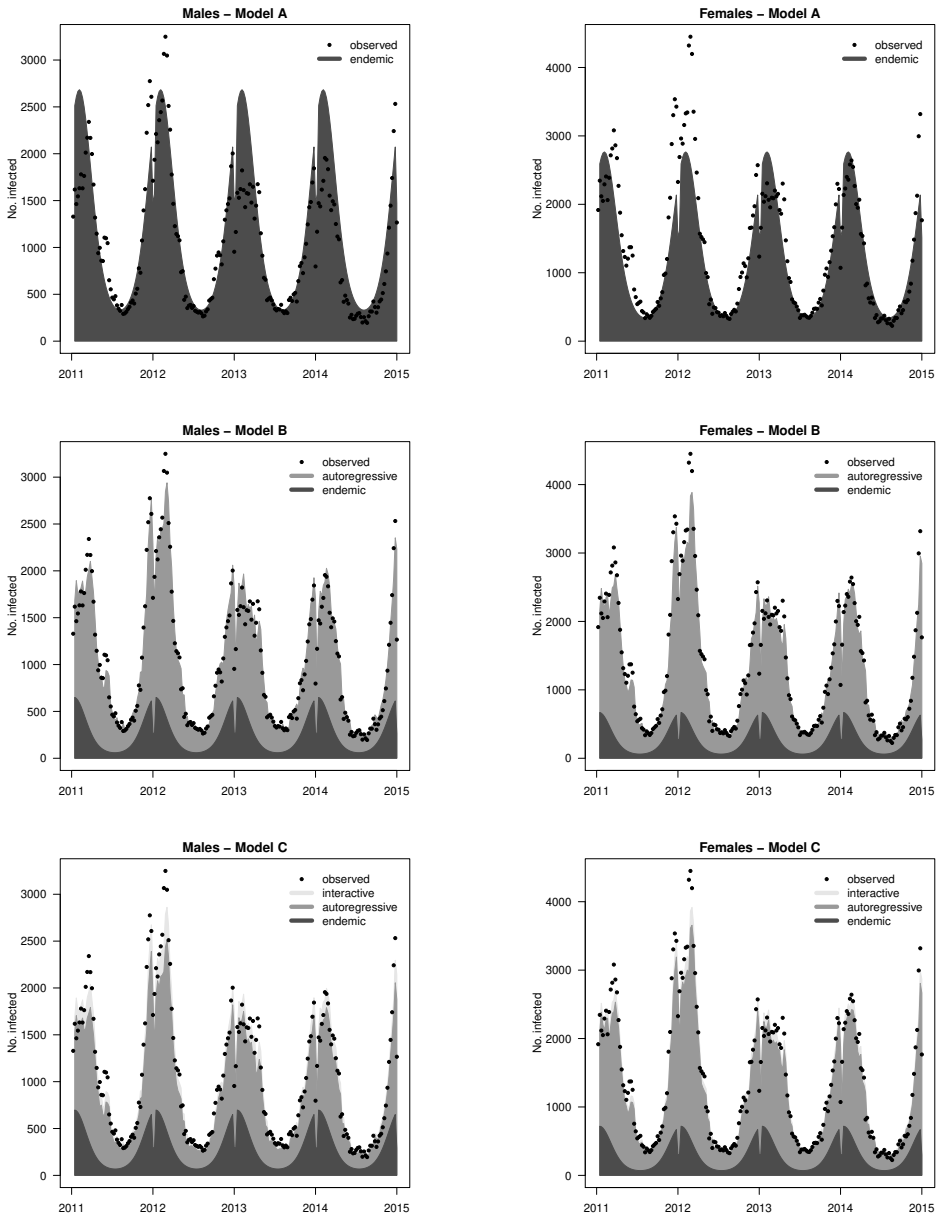


Figure 6: Observed and fitted number of cases in model A–C.

Test	**Model A**		**Model B**		**Model C**	
	p-value	z-value	p-value	z-value	p-value	z-value
Unconditional						
mDSS-1	<0.001		0.006		0.21	
mDSS-2	<0.001	20.05	0.007	2.67	0.42	0.80
iDSS-App.1	0.019		0.97		0.95	
Males	0.69	-0.40	0.95	-0.06	0.93	-0.09
Females	0.004	-2.87	0.82	-0.23	0.75	-0.32
iDSS-App.2	<0.001		0.043		0.084	
iDSS-App.3	0.095	-1.67	0.88	-0.15	0.83	-0.22
Regression						
mDSS	<0.001		0.19		0.66	
iDSS-App.1	<0.001		0.92		0.93	
Males	0.12		0.86		0.92	
Females	<0.001		0.74		0.70	
iDSS-App.2	0.003		0.86		0.91	
iDSS-App.3	0.003		0.77		0.80	

Table 5: The p-values and z-values (if available) from the calibration tests for Model A–C incorporating the correlation coefficient $\hat{\rho}$ of the residuals: mDSS-1 is the test of mDSS based on χ^2-distribution, mDSS-2 is the unconditional test of mDSS, App.1–3 refer to tests of Approach 1–3 based on iDSS.

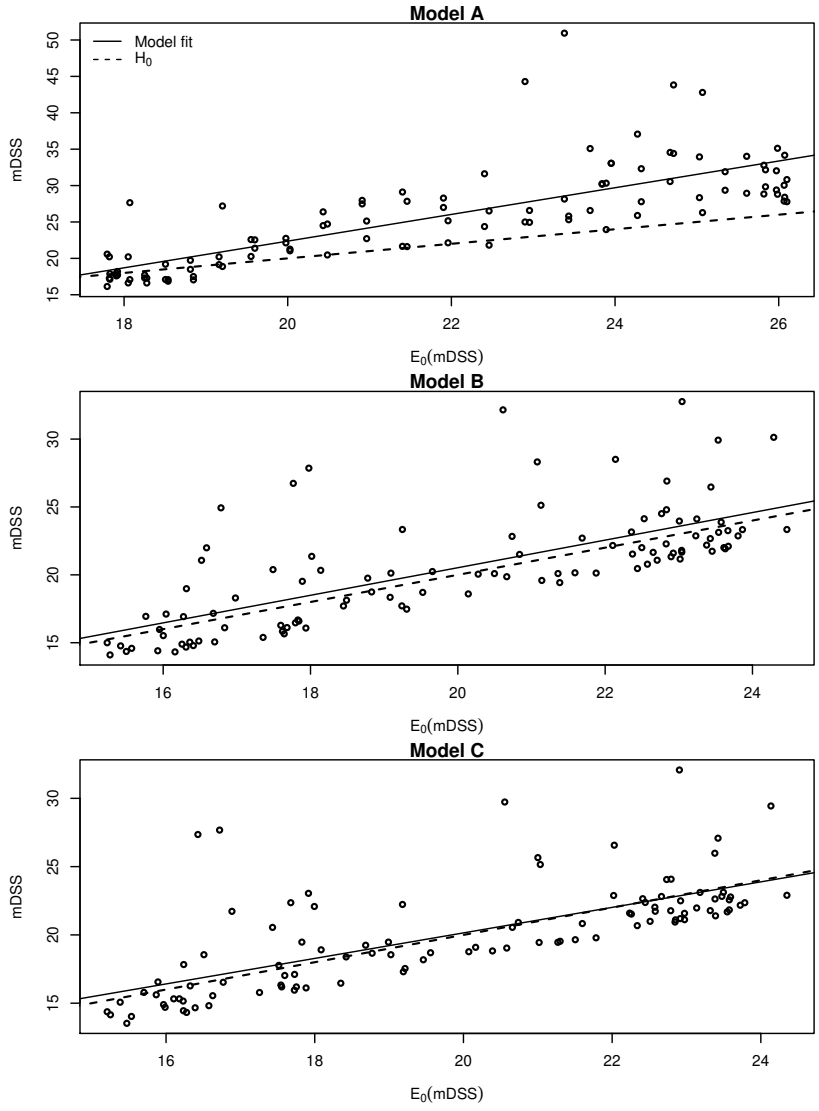


Figure 7: Regression of mDSS on E_0(mDSS) for Model A–C

the unconditional tests and regression tests based on two types of Dawid-Sebastiani scores: mDSS and iDSS. The mDSS gives a scalar value based on multivariate prediction and observation, while iDSS is based on the individual components of the observation and the marginal distributions of the multivariate forecast. We obtained the covariance structure of iDSS and further proposed calibration tests based on either the covariance structure or the multiplicity adjustment (assuming independence of iDSS) using Fisher's method. Dependence among statistical tests based on iDSS is positive, so that the statistic R, which combines the p-values, does not follow a χ^2-distribution. Thus, if Fisher's method for independent tests is applied in a dependent setting, the result is not reliable. A possible extension of Fisher's method can be applied by including the correlation structure, for example Brown's method [7]. Alternatively, approaches developed to address the multiple testing problem can be applied, among which the method developed by Benjamini and Hochberg [4] to control the false discovery rate can be a solution especially for forecasts with high dimensions.

We further evaluated the performance of each test. For the test of mDSS based on the χ^2-distribution, we provided the analytic formula for power calculation. Moreover, this test is powerful even if the number of observations is low because it is not based on the central limit theorem. When the number of observations is large, the unconditional tests based on mDSS are recommended, whose asymptotic power converges to 100%. Further simulation results have shown the other tests proposed are powerful tools for detecting forecasts with either incorrect mean parameters or incorrect variances. Unfortunately, all tests have low power in scenarios where the predictions are incorrectly assumed to be independent. We have also checked the performance of each test when the true generation process is not from a Gaussian distribution via simulations. These show that all tests have sufficient power to detect miscalibration. The application to *Norovirus* disease incidence data from Germany has illustrated that the calibration tests are a useful tool to

detect miscalibration of forecasts for large count data.

In empirical practice, the correlation coefficients among the components are assumed to be fixed. For example, in weather forecast, the daily maximum and minimum temperatures are positively correlated and this correlation is supposed not to change from today to tomorrow [32]; in a time series analysis of exchange rates among different currencies, the correlation is assumed to be invariant with time [37]. Therefore, we have assumed the correlations are fixed in all calibration tests. If varying correlation is desired, the calibration tests discussed can be extended to evaluate this type of forecasts. The tests based on mDSS are valid for this type of forecasts, however we need to modify the tests to include this varying correlation structure of iDSSs.

Among the tests proposed in this paper, the tests based on mDSS have shown superior results to those based on iDSS. We have also applied other tests based on iDSS, for example a test based on the first order statistic (the maximum) of iDSS [11] and found that the results are slightly worse in power compared to the unconditional tests based on iDSS with *Approaches 1–3*. For a multivariate model evaluation, if a model assumes independence among the variables, mDSS and iDSS work equally well when the number of observations is sufficiently large. However, if any correlation structure is included in the modelling, we recommend the tests based on mDSS, which are more powerful even with fewer observations.

Acknowledgments

Financial support by the Swiss National Science Foundation (SNF) is gratefully acknowledged [project #137919]. We acknowledge helpful comments from two referees and thank Isaac Gravestock for useful suggestions to polish the style of the manuscript.

Appendix

Appendix A: Proof of Equation (2.2)

Any two components of a multivariate normal distribution are bivariate normally distributed:

$$\begin{pmatrix} X_j \\ X_k \end{pmatrix} \sim \boldsymbol{\mathcal{N}}_2 \left(\begin{pmatrix} \mu_j \\ \mu_k \end{pmatrix}, \begin{pmatrix} \sigma_j^2 & \rho_{jk}\sigma_j\sigma_k \\ \rho_{jk}\sigma_j\sigma_k & \sigma_k^2 \end{pmatrix} \right).$$

With $\mathrm{iDSS}(x_j, P_j) = \ln \sigma_j^2 + (x_j - \mu_j)^2/\sigma_j^2$ and $\mathsf{E}_0\{\mathrm{iDSS}(x_j, P_j)\} = \ln \sigma_j^2 + 1$,

$$\begin{aligned}
&\mathrm{Cov}_0\{\mathrm{iDSS}(X_j, P_j), \mathrm{iDSS}(X_k, P_k)\} \\
&= \mathsf{E}_0\{\mathrm{iDSS}(X_j, P_j) - \mathsf{E}_0(\mathrm{iDSS}(X_j, P_j))\}\{\mathrm{iDSS}(X_k, P_k) - \mathsf{E}_0(\mathrm{iDSS}(X_k, P_k))\} \\
&= \mathsf{E}_0 \left\{ \frac{(X_j - \mu_j)^2}{\sigma_j^2} - 1 \right\} \left\{ \frac{(X_k - \mu_k)^2}{\sigma_k^2} - 1 \right\} \\
&=: \mathsf{E}_0(Y_j^2 - 1)(Y_k^2 - 1),
\end{aligned}$$

where $Y_j = (X_j - \mu_j)/\sigma_j$ and

$$\begin{pmatrix} Y_j \\ Y_k \end{pmatrix} \sim \boldsymbol{\mathcal{N}}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{jk} \\ \rho_{jk} & 1 \end{pmatrix} \right).$$

With Isserlis' theorem [25] we can show that $\mathsf{E}_0(Y_j^2 Y_k^2) = 1 + 2\rho_{jk}^2$ and obtain

$$\mathrm{Cov}_0\{\mathrm{iDSS}(X_j, P_j), \mathrm{iDSS}(X_k, P_k)\} = 2\rho_{jk}^2,$$

and with $\mathsf{Var}_0(\mathrm{iDSS}(X_j, P_j)) = 2$ we finally have

$$\mathrm{Corr}_0\{\mathrm{iDSS}(X_j, P_j), \mathrm{iDSS}(X_k, P_k)\} = \rho_{jk}^2.$$

Appendix B: Proof of Theorem 4.1

In the mDSS formula, we are interested in the quadratic form

$$Q(\boldsymbol{X}, \boldsymbol{P}) = (\boldsymbol{X} - \boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_P).$$

Since the covariance matrix $\mathbf{\Sigma}_X$ is positive definite, we can find a non-singular matrix $\mathbf{\Gamma}_X = \mathbf{\Sigma}_X^{1/2}$ such that $\mathbf{\Sigma}_X = \mathbf{\Gamma}_X \mathbf{\Gamma}_X^\top$. To make the notation less cumbersome, we will write $\mathbf{\Gamma}$ for $\mathbf{\Gamma}_X$. Let $\boldsymbol{z} = \mathbf{\Gamma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X)$. It is clear that $\boldsymbol{z}$ is distributed according to a standard m-dimensional normal distribution. Since $\mathbf{\Gamma}^\top \mathbf{\Sigma}_P^{-1} \mathbf{\Gamma}$ is a symmetric matrix, we can find an orthogonal matrix $\boldsymbol{L}$ and a diagonal matrix $\mathbf{\Lambda}$ such that $\mathbf{\Gamma}^\top \mathbf{\Sigma}_P^{-1} \mathbf{\Gamma} = \boldsymbol{L}^\top \mathbf{\Lambda} \boldsymbol{L}$. With $\boldsymbol{a} = \mathbf{\Gamma}^{-1}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_P)$ we can write

$$\begin{aligned} Q(\boldsymbol{X}, \boldsymbol{P}) &= (\boldsymbol{X} - \boldsymbol{\mu}_P)^\top \mathbf{\Sigma}_P^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_P) \\ &= (\boldsymbol{z} + \mathbf{\Gamma}^{-1}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_P))^\top \mathbf{\Gamma}^\top \mathbf{\Sigma}_P^{-1} \mathbf{\Gamma} (\boldsymbol{z} + \mathbf{\Gamma}^{-1}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_P)) \\ &= (\boldsymbol{z} + \boldsymbol{a})^\top \boldsymbol{L}^\top \mathbf{\Lambda} \boldsymbol{L} (\boldsymbol{z} + \boldsymbol{a}) \\ &= (\boldsymbol{u} + \boldsymbol{b})^\top \mathbf{\Lambda} (\boldsymbol{u} + \boldsymbol{b}) \end{aligned}$$

where $\boldsymbol{u} = \boldsymbol{L}\boldsymbol{z}$ follows again a standard m-dimensional normal distribution, and $\boldsymbol{b} = \boldsymbol{L}\mathbf{\Gamma}^{-1}(\boldsymbol{\mu}_X - \boldsymbol{\mu}_P)$. It follows that

$$Q(\boldsymbol{X}, \boldsymbol{P}) = \sum_{j=1}^{m} \lambda_j (u_j + b_j)^2$$

where $(u_j + b_j)^2, j = 1, \ldots, m$ are independent random variables such that $(u_j + b_j)^2 \sim \chi^2(1, \nu_j)$ with non-centrality parameter $\nu_j = b_j^2$.

Appendix C

Proposition. *Let $m \in \mathbb{R}$ and $v \in (0, \infty)$. If $W_1, \ldots, W_n$ are i.i.d. random variables with common mean μ and variance $\sigma^2 < \infty$, then, as $n \to \infty$,*

$$\frac{\sqrt{n}\left|\bar{W}_n - m\right|}{v} \to_d \begin{cases} \left|\mathcal{N}(0, \sigma^2/v^2)\right|, & \text{if } \mu = m \\ \infty, & \text{otherwise.} \end{cases}$$

Proof. We have

$$\frac{\sqrt{n}\left(\bar{W}_n - m\right)}{v} = \frac{\sqrt{n}\left(\bar{W}_n - \mu\right)}{\sigma} \frac{\sigma}{v} + \frac{\sqrt{n}(\mu - m)}{v}$$

and the result follows by applying the central limit theorem to $W_1, \ldots, W_n$. □

References

[1] Bao, Y., Lee, T.H., Saltoğlu, B., 2007. Comparing density forecast models. Journal of Forecasting 26, 203–225.

[2] Bartko, J.J., 1966. Approximating the negative binomial. Technometrics 8, 345–350.

[3] Bellman, R.E., 2003. Dynamic Programming. Dover Publications, Incorporated.

[4] Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.

[5] Bernard, H., Werber, D., Höhle, M., 2014. Estimating the underreporting of *Norovirus* illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing E. coli O104:H4 in 2011 – a time series analysis. BMC Infectious Diseases 14, 116.

[6] Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3.

[7] Brown, M.B., 1975. 400: A method for combining non-independent, one-sided tests of significance. Biometrics 31, 987–992.

[8] Statistisches Bundesamt, 2014. Bevölkerung auf Grundlage des Zensus 2011.

[9] Clements, M.P., 2005. Evaluating Econometric Forecasts of Economic and Financial Variables. Palgrave Macmillan.

[10] Cox, D.R., 1958. Two further applications of a model for binary regression. Biometrika 45, 562–565.

[11] David, H.A., Nagaraja, H.N., 2006. Order Statistics, in: Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc.

[12] Dawid, A.P., 1984. Statistical theory: the prequential appoach. Journal of the Royal Statistical Society, Series A 147, 278–292.

[13] Dawid, A.P., Sebastiani, P., 1999. Coherent dispersion criteria for optimal experimental design. Annals of Statistics 27, 65–81.

[14] Diebold, F.X., Gunther, T.A., Tay, A.S., 1998. Evaluating density forecasts with applications to financial risk management. International Economic Review 39, pp. 863–883.

[15] Diebold, F.X., Hahn, J., Tay, A.S., 1999. Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. Review of Economics and Statistics 81, 661–673.

[16] Diggle, P.J., Heagerty, P.J., Liang, K.Y., Zeger, S.L., 2003. Analysis of Longitudinal Data. Oxford Statistical Science Series, Oxford University Press, Oxford.

[17] Fisher, R.A., 1958. Statistical Methods for Research Workers. 13th ed.(rev.) ed., Oliver & Boyd, Edinburgh.

[18] Fretz, R., Svoboda, P., Lüthi, T., Tanner, M., Baumgartner, A., 2005. Outbreaks of gastroenteritis due to infections with *Norovirus* in Switzerland, 2001–2003. Epidemiology and Infection 133, 429–437.

[19] Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 69, 243–268.

[20] Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annual Review of Statistics and Its Application 1, 125–151.

[21] Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L., Johnson, N.A., 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. Test 17, 211–235.

[22] Good, I.J., 1952. Rational decisions. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 14, 107–114.

[23] Held, L., Rufibach, K., Balabdaoui, F., 2010. A score regression approach to assess calibration of continuous probabilistic predictions. Biometrics 66, 1295–1305.

[24] Hotelling, H., 1931. The generalization of Student's ratio. Annals of Mathematical Statistics 2, 360–378.

[25] Isserlis, L., 1918. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. Biometrika 12, 134–139.

[26] Jolliffe, I.T., Stephenson, D.B., 2012. Forecast Verification: a Practitioner's Guide in Atmospheric Science. John Wiley & Sons.

[27] Lopman, B., Vennema, H., Kohli, E., Pothier, P., Sanchez, A., Negredo, A., Buesa, J., Schreier, E., Gray, J., Gallimore, C., et al., 2004. Increase in viral gastroenteritis outbreaks in Europe and epidemic spread of new *Norovirus* variant. The Lancet 363, 682–688.

[28] Lumley, T., 2014. Survey: Analysis of complex survey samples. R package version 3.30.

[29] Mallik, R.K., 2007. Some properties of the uniform correlation matrix and their applications, in: Wireless Communications and Networking Conference (WCNC), 2007, IEEE, pp. 1052–1057.

[30] Mason, S.J., Galpin, J.S., Goddard, L., Graham, N.E., Rajartnam, B., 2007. Conditional exceedance probabilities. Monthly Weather Review 135, 363–372.

[31] Meyer, S., Held, L., Höhle, M., 2016. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. Journal of Statistical Software Preprint available at `http://arxiv.org/abs/1411.0416`.

[32] Möller, A., Lenkoski, A., Thorarinsdottir, T.L., 2013. Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. Quarterly Journal of the Royal Meteorological Society 139, 981–991.

[33] Paul, M., Held, L., 2011. Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. Statistics in Medicine 30, 1118–1136.

[34] Paul, M., Held, L., Toschke, A., 2008. Multivariate modelling of infectious disease surveillance data. Statistics in Medicine 27, 6250–6267.

[35] Riebler, A., Held, L., 2017. Projecting the future burden of cancer: Bayesian age-period-cohort analysis ready for routine use. Biometrical Journal To appear.

[36] Scheuerer, M., Hamill, T.M., 2015. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. Monthly Weather Review 143, 1321–1334.

[37] Sihabuddin, A., Subanar, D.R., Winarko, E., 2014. A second correlation method for multivariate exchange rates forecasting. International Journal of Advanced Computer Science and Applications 5, 30–33.

[38] Spiegelhalter, D.J., 1986. Probabilistic prediction in patient management and clinical trials. Statistics in Medicine 5, 421–433.

[39] Thorarinsdottir, T.L., Scheuerer, M., Heinz, C., 2016. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. Journal of Computational and Graphical Statistics 25, 105–122.

[40] Wei, W., Held, L., 2014. Calibration tests for count data. Test 23, 787–805.

[41] Wu, H.M., Fornek, M., Schwab, K.J., Chapin, A.R., Gibson, K., Schwab, E., Spencer, C., Henning, K., 2005. A *Norovirus* outbreak at a long-term-care facility: the role of environmental surface contamination. Infection Control & Hospital Epidemiology 26, 802–810.